Sharing and Accessing Biodiversity Data Globally through GBIF

Hannu Saarenmaa

Abstract

The Global Biodiversity Information Facility (GBIF) opened its information system in early 2004. GBIF's data portal now integrates tens of millions of records of primary biodiversity data from hundreds of databases worldwide in museums, botanical gardens, and observation networks such as those of bird watchers. This presentation will cover how one can become a GBIF data provider, and how users can access the data using web services and the GIS functions on the GBIF data portal at http://www.gbif.net.

1. Primary Biodiversity Data and Its Uses – an Introduction

Making biodiversity data openly available on the Internet is the goal of the Global Biodiversity Information Facility (GBIF). It was started as the result of an OECD megascience initiative, but now is an independent international organization whose members are 47 countries and 30 other international organizations. This paper gives an overview of GBIF and in particular describes the current processing and future directions for geographic information in the GBIF network.

In the early phases of its life, GBIF has been concentrating on what is called "primary" biodiversity data. Primary biodiversity data originates from field observations of people such as bird watchers, vouchered specimens such as those of botanical and zoological museums, systematic surveys such as natural resource inventories, and other similar sources. The relationships of primary data with other types of data can be illustrated by the pyramid of information (Figure 1).

Because the databases of the data providers include features such as scientific names of organisms and geographic attributes in common, the data can be integrated in different ways. Primary biodiversity data thus represents a huge reuse value. It can be aggregated into distribution maps, provide retrospective views of species distributions, be projected into the future based on various environmental change scenarios, etc. Advanced techniques for doing these kinds of analyses are available, for instance using the GARP tool for ecological niche modeling (http://www.lifemapper.org/desktopgarp/). Such analyses have been used for projecting species invasions, designing reintroduction programs, understanding the effects of global climate and other types of change, understanding rare and endangered species' distributions, and designing biodiversity conservation plans (Peterson 2001, Peterson & al. 2002, Sobéron & Peterson 2004). If data sharing through GBIF keeps growing at its current rate, it will be very useful for

answering the big burning questions about loss of biodiversity and help inform international agreements on reversing such trends by the year 2010 (http://www.biodiv.org/2010-target/).

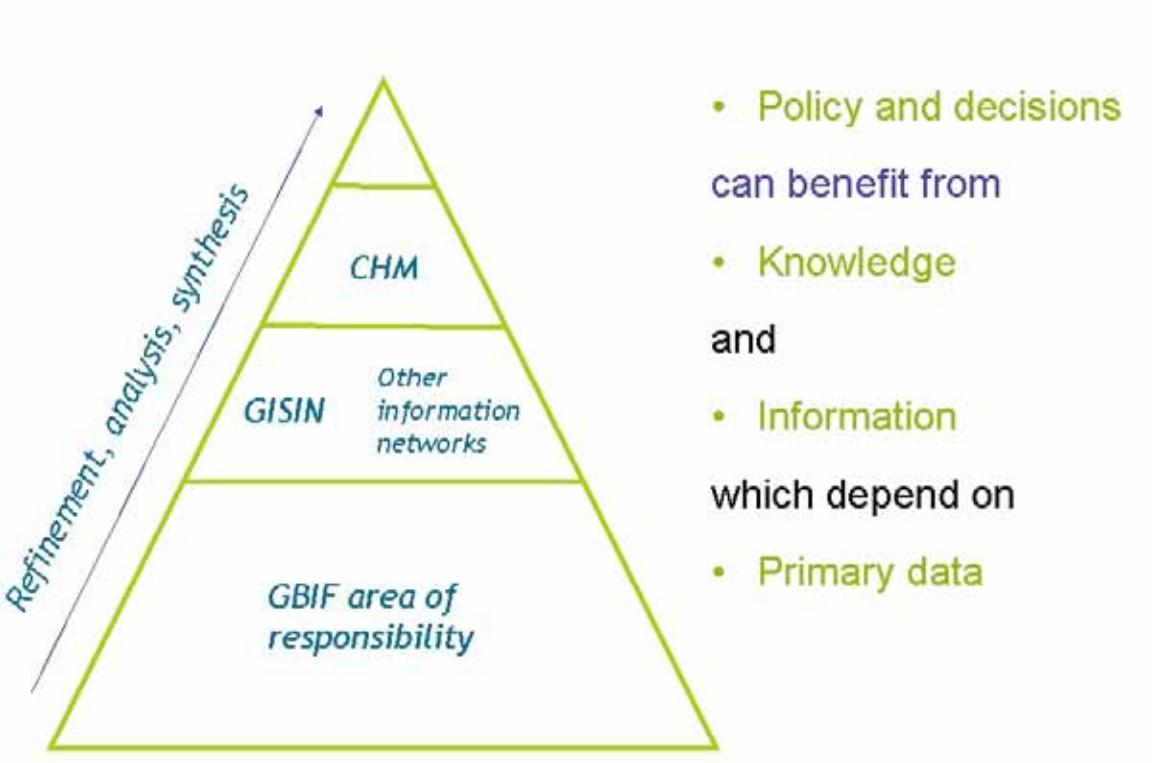

Figure 1. GBIF is currently focusing on primary biodiversity data at the bottom of the "pyramid of information".

2. Current Status of GBIF Data

The GBIF information system became on-line in early 2004 and after one and a half years of operation, 73 million records of primary biodiversity data from 521 collections/databases have been made available on 127 data providers by 35 GBIF Participants. This data has been collected from 239 countries or territories. GBIF estimates that these figures only represent about 20% of existing digitized biodiversity data. That is, data served via GBIF still is quite patchy, but the gaps are being filled rapidly. The biggest gains of data have been achieved by linking with GBIF on-line observer systems such as Cornell Laboratory of Ornithology, the UK's National Biodiversity Network, and the Swedish Species Gateway, and also by connecting to existing thematic networks such as OBIS and MaNIS, (see http://www.gbif.org/DataProviders/).

The data providers share their data using standard protocols and standard data exchange formats. For the latter, two XML schemas are currently being used, which are ABCD

(http://www.bgbm.org/tdwg/CODATA/Schema/) and Darwin Core (http://darwincore.calacademy.org/). The former offers a complete metamodel of a biological collection with several hundreds of nested elements. The latter is a minimalistic flat schema with 48 elements.

Naturally, both schemas include latitude and longitude, and among the current GBIF data, 74% actually does have values for these. Indicating coordinate precision is also available, but only 10% of the current data offers some value for this.

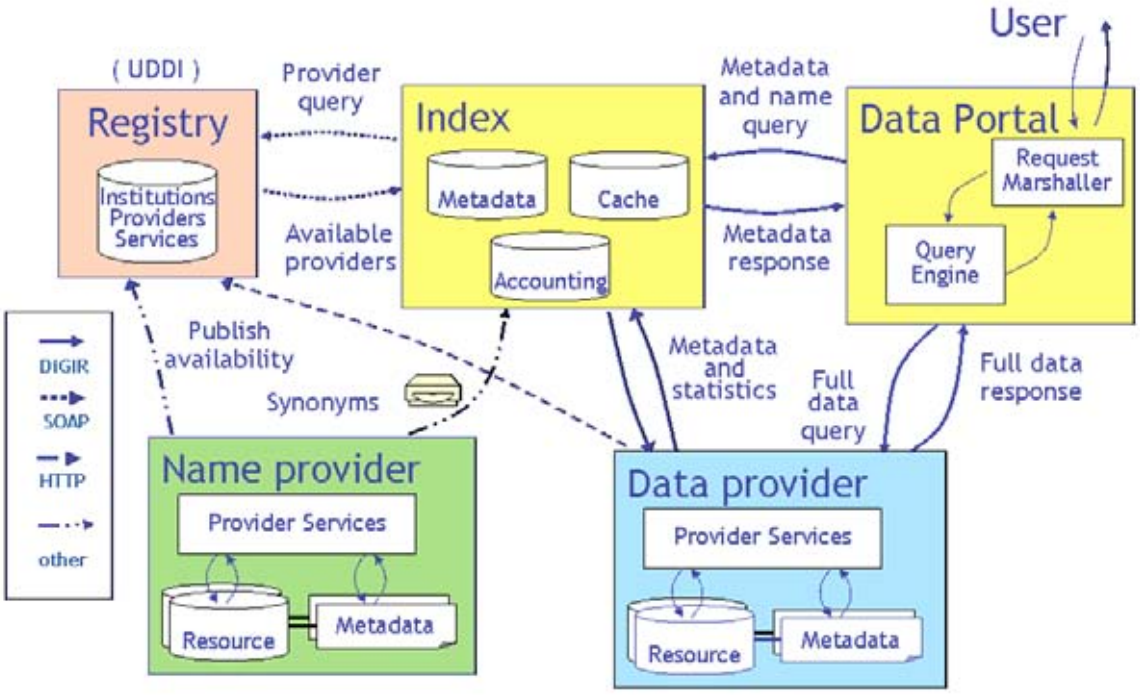

Figure 2. Component architecture of the GBIF information system.

3. Architecture and Components of the GBIF Information System

The GBIF information system is based on a web services approach. This means that there are distributed data providers, a central registry of them, and a central portal to access the data (Figure 2), all communicating using standardized XML messages. This standardization is done through the Taxonomic Databases Working Group (http://www.tdwg.org/)

The **data providers** must install wrapper software that maps the local database schema into the ABCD or Darwin Core formats, translates XML-encoded queries coming from internet into SQL, and return data or metadata the same way. The two currently supported protocols for this communication are BioCASe (http://www.biocase.org/ and

DiGIR (http://www.digir.net/). Work is underway to build a successor for these, called TAPIR (http://ww3.bgbm.org/protocolwiki/), which merges the best parts of each protocol. SOAP is not being used here (yet) because it would introduce extra overhead for moving around large datasets.

The data providers can announce their presence in the GBIF **UDDI registry** (http://registry.gbif.net/). This is a central service that is based on a commercial tool from Systinet, Inc. Anyone can install the data provider software and register. However, after registration, a Node Manager from an existing GBIF participant (coordinator from a country/ or organization) must endorse the data provider as somebody that is indeed sharing scientific biodiversity data. This is a rudimentary quality assurance step, although the Node Manager does not currently scrutinize the actual data records. Before such an endorsement is received, GBIF does not make the data available, although the UDDI registry has open SOAP interfaces for other portals to discover the data providers.

GBIF operates a prototype **data portal** at http://www.gbif.net/ that can be used to search, browse and drill into the data of the endorsed data providers. This central gateway to GBIF data is multilingual and it maintains a central index of the most important data elements of all the records in the data providers. There also is a synonym resolution of organism names provided by the Catalogue of Life Partnership (http://www.sp2000.org/), although currently only about 30% of the species can be covered this way. The data portal currently does not offer XML data services, but as high performance TAPIR implementations become available these functionalities will be added.

4. Geographic Information Processing at GBIF

The GBIF data portal currently has no geographic query capabilities, but it does offer links to some geographic visualization services provided by GBIF participants.

All results of queries are sent to the Belgian Biodiversity Information Facility, which results in a simple GIF overview map of the distribution of the data points. This service is based on MapServer and PostGIS.

There also are two levels of dynamic map service from the Canadian Biodiversity Information Facility (CBIF). Upon request by the user, the data are sent from the GBIF data portal to a map server in Canada. In the original implementation, this service offers zooming, panning, and drilling into the entire records by clicking the dots shown on a world map layer from Demis, a company based in the Netherlands (http://www2.demis.nl). The "Help on Icons" link below the map explains the use of the various buttons on the left side of the map.

Recently, more functionality has been achieved through an experimental custom interface to GBIF's global index. This server at http://www.cbif.gc.ca/mapdata/cbif/ offers a Web Map Service (WMS) access to these records. The WMS service makes dynamic queries against the GBIF data portal based on scientific names input by the user. The requirement for scientific names might inhibit some users, but one can search the GBIF

names data to find the scientific name that corresponds to a common name. Multiple species layers can be plotted together on one map, using differing shapes, colors and sizes of points selected by the user. Each of the plotted coordinates is queriable and will return the user to the GBIF data portal to retrieve the information for a specific point (using the GetFeatureInfo query).

When a user visits the CBIF portal and clicks on "Online Mapping", he or she is taken to a page that has three links. The first of these allows a user who does not have any specialized geographic information systems (GIS) knowledge to utilize the web mapping service (WMS) that CBIF has created, as described above. The next link leads to the "Generic Point Mapper", which is a tool that anyone can use to generate his or her own dynamic map (similar to the one seen in the "Map GBIF Occurrence Data" link). These interactive maps allow a user to zoom in or out, to reposition the center of the map, and the like. Users can access this tool to build such maps for their own websites. Finally, the actual WMS that underlies the "Map GBIF Occurrence Data" service can be accessed by programmers and developers who do have specialized GIS knowledge and wish to add GBIF species occurrence layers to other online GIS applications.

Similar mapping services are provided by other sites, but these do not currently call on the GBIF central data portal for the data used. Berkeley mapper (http://mvz.berkeley.edu/DiGIR_Mapper.html) works against those DiGIR providers that belong to the MaNIS and HerpNet networks. CRIA in Campinas, Brazil, offers a wide suite of geographic and other data quality checking services for the Brazilian data providers in the SpeciesLink network (http://splink.cria.org.br/).

5. Becoming a GBIF Data Provider

It is important to understand that GBIF is not just a secretariat based in Copenhagen or just a data portal, but that GBIF is all the institutions, corporations, and individuals whose countries or organizations have signed the Memorandum of Understanding that mandates GBIF. In so doing, they have expressed their willingness to openly share biodiversity data, in the spirit of important initiatives such as the Conservation Commons (http://www.conservationcommons.org/) that highlight the importance of open sharing of data for societal and scientific benefit.

The GBIF data sharing agreement upholds its principles of sharing biodiversity data openly for common benefit. It is only through combined efforts of everyone who own such data that some of the most burning environmental questions of our time can be answered, and new scientific discovery be achieved. Sharing data is a scientific responsibility all taxonomists, observer networks, and environmental surveys.

Becoming a GBIF data provider is easy and in many cases will not take more than a few hours. It includes a few steps that are explained on the GBIF website (http://www.gbif.org/DataProviders/HowTo/). The technical task is downloading and configuring the BioCASe or DiGIR -based data provider software. There are three integrated packages for Windows and Linux that install in mere minutes and are

supported through a central helpdesk and documentation. These packages have been put together from open source components and GBIF naturally provides these packages, as well as central web services and registry, for free, thus enabling and facilitating data holders to share it openly.

6. Future Steps and Conclusions

GBIF will soon be adding a few more data types and going beyond primary data. There will be data provider tools for names and checklists. Species home pages including digital images are mushrooming on the Internet. Linking to these will add value to the shared primary data. Images can be shared by BioCASe/ABCD providers, but extending that capability to all data providers is an important extension to the Darwin Core currently under development.

Interfaces for accessing GBIF portals and data providers directly from GIS software tools will be also soon be available. The new TAPIR provider will have an OGC Web Feature Service. These interfaces can be registered in the UDDI registry as additional alternate bindings.

Recently, GBIF has experimented with slicing its global index by country. For instance, a slice on all data from Madagascar was used for a geographic demonstrator put together by Conservation International.

GBIF is in the process of installing mirror sites for its central data portal in Germany, Korea, and the United States. There probably also will be specialized service providers that make use of the interfaces of these mirror sites. The experience of the CBIF WMS site has been so good that other similar services can now be encouraged, and standard mechanisms for building such value adding services will have to be established. The data quality assurance services of CRIA and SpeciesLink are an example of what can achieved. Under an agreement between CRIA and GBIF, some of these services are being fitted also for the GBIF data portal.

In closing, it should be emphasized that the GBIF data portal does not intend to be all things for all users. Anyone can build portals that access GBIF registered data providers, because this is technically enabled by the open interfaces of the GBIF UDDI registry. Some specialized communities have already taken advantage of this, such as the Lifemapper project (http://www.lifemapper.org/) that made use of this open availability.

References

Peterson, A. T. 2001. Predicting species' geographic distributions based on ecological niche modeling. *Condor*, 103:599-605.

Peterson, A. T., M. A. Ortega-Huerta, J. Bartley, V. Sanchez-Cordero, J. Soberon, R. H. Buddemeier, and D. R. B. Stockwell. 2002. Future projections for Mexican faunas under global climate change scenarios. *Nature*, 416:626-629.

Soberón, J., and A. T. Peterson. 2004. Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London B*, 359:689-698.

Acknowledgments

Guy Baillargeon, Derek Munro and Françoise Guilbault of the Canadian Biodiversity Information Facility built the dynamic web mapping service and contributed text to this article. Johan Duflost, Patricia Mergen and Frédéric Wautelet of the Belgian Biodiversity Information Facility built the Belgian mapping service. Donald Hobern, Meredith Lane, and Larry Speers of GBIF Secretariat contributed text, reviewed this article and improved the language. Thanks to them all.

Author information

Hannu Saarenmaa
Deputy Director for Informatics
GBIF Secretariat
Universitetsparken 15
2100 Copenhagen
Denmark
Telephone +45-35321479
Email hsaarenmaa@gbif.org